Simplify your big data journey with a tested and validated Hadoop solution

Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture

Gaining more value from your enterprise data warehouse

In many large organizations, the enterprise data warehouse has emerged as the central data store for business reporting, data management and data ingestion from diverse sources, both inside and outside the enterprise. In this role as the keeper of all corporate data, the EDW provides a foundation for advanced analytics solutions and a path to the data-driven insights that guide the future-ready enterprise.

While the EDW plays an all-important role in the effort to leverage big data to drive business value, it is not without its challenges. In particular, the typical EDW is being pushed to its limits by the volume, velocity and variety of data. According to Gartner, 70 percent of EDWs are performance- and capacity-constrained. In particular, data warehouses are sagging under the weight of extract/transform/load (ETL) processes, which take raw data from source systems, manipulate it into a consumable format, and load it into a system that handles advanced analytics and reporting.

To complicate the challenge, diverse data formats—including new unstructured and semi-structured data types—are adding ETL complexity and creating even heavier processing burdens. By some estimates, data integration and transformation workloads can now consume as much as 80 percent of EDW capacity, creating bottlenecks in the EDW. In many cases, just a few heavy jobs can bog down an EDW, threatening service level agreements (SLAs) with business units.

These backend challenges driven by an overloaded EDW can impact the frontlines of the business, because more ETL processing means less query capacity. For example, business segments might not be able to run crucial reports in time to make critical business decisions, and business analysts might not be able to query data for ad hoc analysis, resulting in slower decision processes. This is a problem that isn't going to go away. It's only going to grow worse as the universe of big data grows in size and complexity.

So how do you overcome these challenges? It might seem that the logical path forward is to scale out your EDW, but that can be a costly proposition in terms of software licensing,

The analyst view

"Dell is one of a very few companies possessing the right ingredients to really reshape the big data and analytics market, including the brand presence, a full hardware stack, a broad range of data-oriented software, and the professional services muscle to make it all come together."

—Nik Rouda, Senior Analyst, ESG

ESG Research Report, Dell's Big Picture for Big Data and Analytics, June 2014.

infrastructure and consulting fees. And even at that, many EDWs are not equipped to handle the diverse variety of today's data—from social media feeds to machine-generated data streams.

Instead, these challenges drive the need for solutions that offload the heavy lifting of ETL processing from the data warehouse to a complementary, lower-cost processing environment built for the diversity of today's data. This approach frees up capacity in your EDW, which can then be devoted to higher-value analytics queries and business reporting—the work that results in business insights.

This is where the Dell™ | Cloudera™ | Syncsort™ Data Warehouse Optimization – ETL Offload Reference Architecture enters the picture.

A proven solution

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture enables your organization to lower data transformation costs and build operational efficiencies while laying a robust, cost-effective, secure and scalable foundation for managing data while maturing into advanced data analytics.

Jointly designed by Dell, Cloudera, Intel and Syncsort, the tested and validated Reference Architecture outlines the end-to-end components for a complete ETL offload solution. The solution includes all the hardware, software, resources and services you need to turn Hadoop into a robust ETL environment. With this industry-first end-to-end approach, you can be in production with Hadoop for ETL offload in a shorter time than would be typically possible with a homegrown solution.

Business-driven benefits

Offloading ETL workloads from your EDW to Hadoop can help your organization:

- Achieve significant improvements in business agility
- Avoid unsustainable EDW upgrade costs just to keep the lights on
- Optimize your EDW by freeing up valuable storage and processing capacity for faster queries and other workloads more suitable for the EDW
- Start building your enterprise data hub (EDH)

Here are some of the specific, quantifiable benefits of the Dell | Cloudera | Syncsort solution:

REDUCE OPERATIONAL COSTS

Estimates from multiple sources indicate managing data in Hadoop can range from $250 to $2,000 per terabyte of data, compared to $20,000 to $100,000 per terabyte for high-end data warehouses.

REDUCE CAPITAL COSTS

Relying on the EDW for heavy data transformations can lead to unsustainable costs and complexity. Data integration and transformation workloads can now consume as much as 80 percent of EDW capacity, according to Gartner. It's no wonder that 70 percent of today's data warehouses are performance- and capacity-constrained, according to the firm.[1] With ETL processes driving the bulk of database workloads, it isn't unusual for organizations to spend upwards of $300,000 per year on additional EDW capacity—just to keep the lights on. The Dell | Cloudera | Syncsort solution can help you avoid these costly EDW upgrades.

IMPROVE EDW PERFORMANCE

Under typical conditions, the top 20 percent of ETL workloads can consume up to 80 percent of an EDW's processing capacity. Moving these workloads to Hadoop can improve data-warehouse performance and make resources more readily available to the business for high-value data analytics queries and reporting.

Validated benefits

Research by Principal Technologies, a technology testing and analysis firm, found that the Dell | Cloudera | Syncsort ETL offload solution can help organizations drive operational efficiency by completing Hadoop ETL jobs faster, simplifying the Hadoop ETL design

1 Gartner. "The State of Data Warehousing in 2014." June 19, 2014.

process, and saving thousands of dollars on Hadoop ETL jobs.

In specific terms, the firm determined:

- A Dell | Cloudera | Syncsort solution for Hadoop fully-implemented by an entry-level employee could reduce administrative costs by 76 percent.[2]
- ETL jobs created by an entry-level technician using the Dell | Cloudera | Syncsort solution for Hadoop ran up to 60 percent faster than a solution created by a Hadoop expert using open source tools.[3]
- A Dell | Cloudera | Syncsort solution for Hadoop enables less-experienced users to develop and deploy Hadoop ETL jobs in less than a week.[4]

Ideal use cases

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload solution delivers a use-case driven Hadoop Reference Architecture to guide your data warehouse optimization efforts. Here are some of the more common uses for the solution.

ETL OFFLOAD

ETL offload moves the heavy lifting of extract/transform/load processes out of your enterprise data warehouse and into a highly scalable Hadoop cluster. Shifting this work into Hadoop helps you lower cost and increase operational efficiency by shortening batch windows with fresher data. This data can then be queried faster because the EDW is no longer bogged down in data transformation jobs.

DATA WAREHOUSE OPTIMIZATION

The Dell | Cloudera | Syncsort solution can be used to augment a traditional relational management database or enterprise data warehouse with Hadoop. In this role, Hadoop acts as a single data hub for all data types.

INTEGRATION WITH THE DATA WAREHOUSE

The Dell | Cloudera | Syncsort solution provides the tools you need to extract, transform and load data into and out of Hadoop using a separate database management system for advanced analytics. With this smarter approach, you can collect, process and integrate more data in less time, while reducing the total cost and complexity of your data integration environment. Better still, transformations are processed on the fly, eliminating the need for costly database staging areas or manually pushing transformations to the database.

HIGH-PERFORMANCE DATA TRANSFORMATIONS

The Dell | Cloudera | Syncsort solution is ideal for high-performance data transformations. The included DMX-h ETL software packages a library of hundreds of smart algorithms to help you deal with the most demanding data integration transformations and functions, including sorting, aggregating, joining, parsing, hashing and pattern-matching.

DATA WAREHOUSE SIMPLIFICATION

Moving problematic, resource intensive or performance challenged workloads to Hadoop can lighten the load on the data warehouse making it easier to manage. Migrated workloads can be refactored to scale with Hadoop as needed.

MAINFRAME DATA INGESTION AND TRANSLATION

Most large organizations still rely on mainframe systems to run core applications, generating massive amounts of transaction data every day. Neglecting this data can result in missed business opportunities. The Dell | Cloudera | Syncsort solution can be configured to read files directly from a mainframe system, parse and transform the data and more. You can do all of this without installing any software on the mainframe and without writing any code.

Accelerate tasks

"Both the Syncsort DMX-h and DIY jobs that we created were able to reformat the mainframe data successfully, but the Syncsort DMX-h job was able to complete the task 17.9 percent faster."

—Principled Technologies

Principled Technologies, "Performance Advantages of Hadoop ETL Offload with the Intel-Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

2 Principled Technologies. "Cost Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

3 Principled Technologies. "Performance Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

4 Principled Technologies. "Design Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

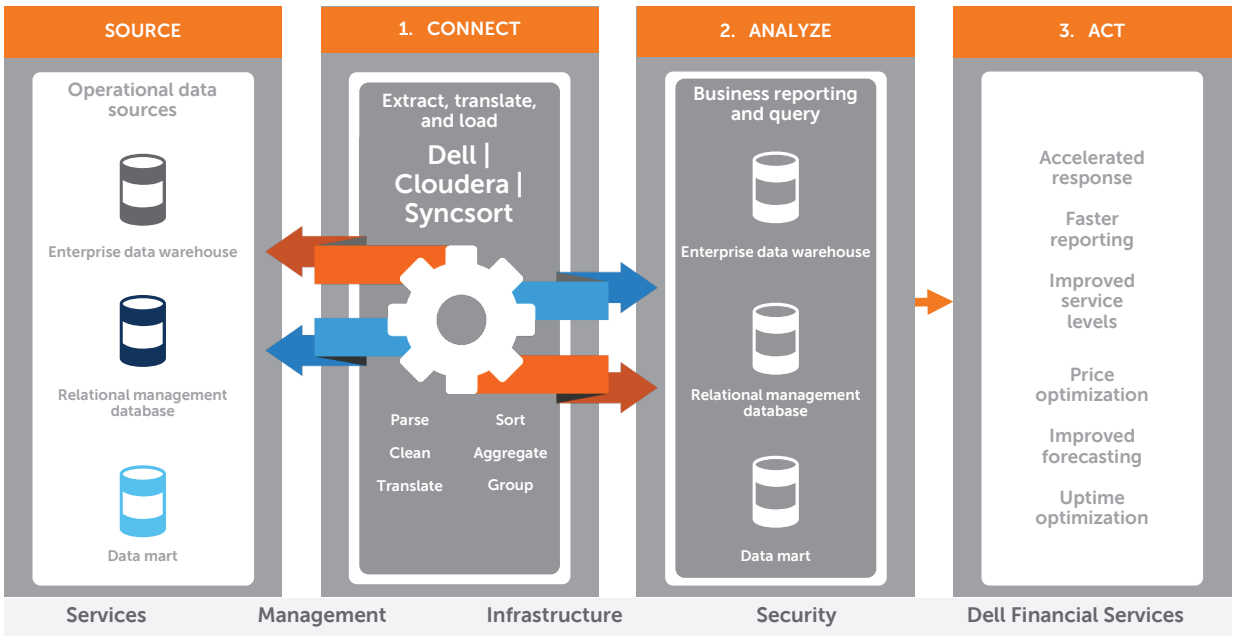

Figure 1: Operational efficiency architecture

Simplify operations

"We found that the Dell | Cloudera | Syncsort ETL offload solution was so easy to use that an entry-level employee could use it to create optimized ETL jobs after only a few days of training."

—Principled Technologies

Principled Technologies, "Design Advantages of Hadoop ETL Offload with the Intel-Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

Solution components

The core components of the solution include the tested and validated reference architecture, software from Cloudera and Syncsort, and hardware and services from Dell. Specific solution components include:

- Cloudera Distribution for Apache Hadoop (CDH) version 5.5.1
- Syncsort DMX-h version 8.5 ETL software
- Dell™ PowerEdge™ R series servers with Intel® Xeon® processors
- Dell™ networking
- Optional Dell consulting and integration services

Let's take a closer look at these components.

CLOUDERA COMPONENTS

Dell's ETL-offload solution delivers the power of the Hadoop environment via Cloudera Enterprise software—the industry-leading distribution of Hadoop. Designed specifically for mission-critical environments, Cloudera Enterprise includes CDH, the world's most popular open source Hadoop-based platform, as well as advanced system and data management tools, open source software components that help with Hadoop usability, and dedicated support from Hadoop experts.

SYNCSORT COMPONENTS

For sophisticated ETL offload capabilities, the solution incorporates Syncsort DMX-h ETL software. This software suite makes it easy, even for non-data-scientists, to build and deploy ETL jobs in Hadoop. With DMX-h, users can start developing Hadoop ETL jobs within hours, and become fully productive within days, using a drag-and-drop interface and the same ETL skills they already have. There's no need to learn complex technologies like MapReduce, Pig or Hive.

To fast-track your EDW-offload projects, the solution includes Syncsort SILQ, a SQL offload utility designed to help users understand and offload complex SQL data integration workloads from a data warehouse into Hadoop. SILQ takes a SQL script as an input and then provides a detailed flow chart of the SQL logic. Using an intuitive web-based interface, you can easily drill down to get detailed information about each step within the data flow, including tables and data transformations.

Improve productivity

"In our tests, we found that the unique design of the Dell | Cloudera | Syncsort ETL offload solution can allow an end user with little experience in using Hadoop to develop and deploy optimized ETL jobs up to 58.8 percent faster than an expert-driven do-it-yourself (DIY) solution deployed using open-source tools."

—Principled Technologies

Principled Technologies, "Design Advantages of Hadoop ETL Offload with the Intel-Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

DELL AND INTEL COMPONENTS

At the hardware level, the solution leverages the Dell PowerEdge R730XD servers. This system delivers the performance, power efficiency, virtualization and security features of the Intel® Xeon® E5 2600 v3 processor, along with large memory capacities and fast storage options.

Other solution components include the Dell S4048-ON Switch Production Network, the Dell S3048-ON Switch Managment and the Dell S6000 Aggregation Switch, along with optional deployment and consulting services.

PROFESSIONAL SERVICES AND SUPPORT

Dell Deployment and Consulting can help you quickly realize the full benefit of your data warehouse optimization investments while limiting business disruptions. Dell can provide on-site deployment of the solution hardware, configuration of servers and network switches, and installation of the solution software. Deployment services are performed by trained experts in accordance with Dell, Cloudera and Syncsort best practices. Additional training and consulting services can be added to help your team complete the offloading of ETL workloads into Hadoop.

Support for the overall solution is provided through Dell ProSupport, with collaborative assistance from the Cloudera and Syncsort support teams.

Are you capitalizing on your data?

Many enterprises are not capitalizing on their data analytics opportunities. According to the Dell Global Technology Adoption Index, 61 percent of organizations studied said they have data that could be analyzed but only 39 percent of organizations understand how to extract value from data and are pursuing that goal.

Nevertheless, awareness is growing around the big data and analytics opportunities. The same Dell Global Technology Adoption Index study found that organizations that have successfully leveraged data over the last three years grew by almost twice as much as those who did not: 14 percent versus 8 percent. The benefits of becoming data-driven are real: The study confirmed that properly leveraging data does lead to competitive advantage.

This idea is at the heart of a future-ready enterprise. To compete effectively in the years to come, enterprises are going to need to take greater advantage of the massive amounts of data they collect. The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture helps you take a step in this direction.

The Reference Architecture covers everything you need to capitalize on your ETL-offload opportunities with an end-to-end solution—including software, hardware and services. With this robust offload solution, your organization can accelerate ETL processing, work easily with a wide range of new data sources and formats, and make better use of existing EDW investments—to generate tremendous business value.

Ultimately, when you work with Dell you gain the confidence that comes with a Reference Architecture based on a proven approach to Hadoop solutions. The Dell | Cloudera | Syncsort ETL offload solution is the 15th Dell Hadoop Reference Architecture that has been certified and validated beginning since 2011. These Reference Architectures are the result of a tested and validated process that Dell has refined over the years to provide the blueprints for an optimal customer experience.

To learn more, visit Dell.com/Hadoop | Dell.com/BigData

© 2016 Dell Inc. All rights reserved. Dell, the DELL logo, the DELL badge and PowerEdge are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

February 2016 | Version 5.5.1